RR-91-24-ONR

NEW VIEWS OF STUDENT LEARNING: IMPLICATIONS FOR EDUCATIONAL MEASUREMENT

Geofferey N. Masters
Australian Council for Educational Research

Robert J. Mislevy
Educational Testing Service

This research was sponsored in part by the
Cognitive Science Program
Cognitive and Neural Sciences Division
Office of Naval Research, under
Contract No. N00014-88-K-0304
R&T 4421552

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, New Jersey

March 1991

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) RR-91-24-ONR	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Educational Testing Service 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Cognitive Science Program, Office of Naval Research (Code 1142CS), 800 North Quincy Street
6c. ADDRESS (City, State, and ZIP Code) Princeton, NJ 08541	7b ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000
8a NAME OF FUNDING / SPONSORING ORGANIZATION 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-88-K-0304
8c. ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61153N	RR04204	RR04204-01	R&T4421552

11 TITLE (Include Security Classification)
New Views of Student Learning: Implications for Educational Measurement (Unclassified)

12 PERSONAL AUTHOR(S)
Geofferey N. Masters, Robert J. Mislevy

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Technical	FROM _____ TO _____	March 1991	41

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	cognitive psychology, test theory, mixed strategies, partial credit model, achievement tests
05	10		

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Recent research in cognitive psychology has drawn attention to the important role that students' personal understandings and representations of subject matter play in the learning process. This paper briefly reviews some of this research, and contrasts the kind of learning that results in an individual's changed conception or view of a phenomenon with the more passive, additive kind of learning assessed by most traditional achievement tests. To be consistent with a view of learning as an active, constructive process,

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Dr. Charles E. Davis	703-696-4046	ONR 1142CS

DD Form 1473, JUN 86 Previous editions are obsolete

S/N 0102-LF-014-6603

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

19 ABSTRACT

educational tests are required which focus on key concepts in an area of learning, and which take into account the variety of types and levels of understanding that students have of those concepts. In these tests, scoring responses right and wrong is likely to be less appropriate than using students' answers to infer their levels of understanding. This will require not only imaginative new types of test items, but statistical models that permit inferences about students' understandings once their responses have been observed. Psychometric approaches are sketched to construct measures of achievement from such tests.

New Views of Student Learning: Implications for Educational Measurement

Geofferey N. Masters

Australian Council for Educational Research

Robert J. Mislevy

Educational Testing Service

March 1991

To appear in N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test Theory for a New Generation of Tests*, published by Lawrence Erlbaum Associates, Hillsdale, NJ. The work of the second author was supported in part by Contract No. N00014-88-K-0304, R&T 4421552, from the Cognitive Sciences Program, Cognitive and Neural Sciences Division, Office of Naval Research; the views expressed herein do not necessarily reflect those of that agency. We are grateful to Isaac Bejar, Norman Frederiksen, Drew Gitomer, and Kikumi Tatsuoka for comments on earlier versions of the paper.

ABSTRACT

Recent research in cognitive psychology has drawn attention to the important role that students' personal understandings and representations of subject matter play in the learning process. This paper briefly reviews some of this research, and contrasts the kind of learning that results in an individual's changed conception or view of a phenomenon with the more passive, additive kind of learning assessed by most traditional achievement tests. To be consistent with a view of learning as an active, constructive process, educational tests are required which focus on key concepts in an area of learning, and which take into account the variety of types and levels of understanding that students have of those concepts. In these tests, scoring responses right and wrong is likely to be less appropriate than using students' answers to infer their levels of understanding. This will require not only imaginative new types of test items, but statistical models that permit inferences about students' understandings once their responses have been observed. Psychometric approaches are sketched to construct measures of achievement from such tests.

1. INTRODUCTION

Implicit in much of our current measurement theory and practice is a view of learners as passive absorbers of provided wisdom. Most items on standard achievement tests assess students' abilities to recall and apply facts and routines presented during instruction. Some require only the memorization of detail; they seek evidence that students have absorbed factual details presented in class and are able to reproduce these on command. Other achievement test items, although supposed to assess higher-level learning outcomes like "comprehension" and "application", often require little more than the ability to recall a formula (e.g., $s = v_0t + 1/2\ at^2$) and to make appropriate substitutions to arrive at a correct answer.

Test items of this type are consistent with a view of learning as a passive, receptive process through which new facts and skills are added to a learner's repertoire in much the same way as bricks might progressively be added to a wall. The process is additive and incremental: students with the highest levels of achievement in an area are those who have absorbed and can reproduce the greatest numbers of facts and formulae. The practice of scoring answers to items of this type either "right" or "wrong" is consistent with the view that individual units of knowledge or skill are either present or absent in a learner at the time of testing. Under this approach, diagnosis is a simple matter of identifying unexpected holes or gaps in a student's store of knowledge. These are subareas of learning in which knowledge is "missing" and in which there is a need for remedial teaching to fill a deficit.

This approach to the measurement of achievement may be appropriate for some forms of learning—as when the learner's task is in fact to master a body of factual material. In recent decades, however, significant advances have occurred in our understanding of the ways in which students learn. In particular, there has been an increased awareness of the active, constructive nature of most forms of human learning and of the important role that students' personal conceptions and representations of subject matter play in the learning process. Rather than being a passive process of absorbing new material as it is

encountered, meaningful learning is increasingly being recognized as an active process through which students construct their own interpretations, approaches, and ways of viewing phenomena, and through which learners relate new information to their existing knowledge and understandings. Under this view of learning, the difference between beginning and advanced learners is seen not so much as a difference in amount of factual knowledge (although this is usually an important aspect of competent performance), as a difference in the types of conceptions and understandings that students bring to a problem, and in the strategies and approaches that they use.

Support for this view of learning can be found in recent studies in a number of areas of investigation. In cognitive science, comparisons of novices and experts in various fields of learning show that expertise typically involves much more than mastery of a body of facts: experts and novices usually have very different ways of viewing phenomena and of representing and approaching problems in a field (e.g., Chi, Feltovich, & Glaser, 1981, in physics; Chase & Simon, 1973, in chess; Lesgold, Feltovich, Glaser, & Wang, 1981, in radiology; and Voss, Greene, Post, & Penner, 1983, in social science). Expert-novice studies suggest that the performances of beginning learners often can be understood in terms of the inappropriate or inefficient models that these learners have constructed for themselves.

Similar observations have been made in the field of science education (see Driver & Easley, 1978; Osborne & Wittrock, 1983; Posner, Strike, Hewson, & Gertzog, 1982). Research into students' science learning has drawn attention to the frequent mismatch between intuitive understandings that students bring to the classroom and the conceptual frameworks assumed by teachers. Caramazza, McCloskey, and Green (1981) observe that the scientific "principles" that students abstract from everyday experience are often strikingly at variance with the most fundamental physical laws. These misunderstandings can go undetected by teachers if correct answers to test questions depend only on superficial knowledge of formulae and formula manipulation techniques (Clement, 1982).

There is evidence that students can succeed in high school and even college science courses while still maintaining many of their misconceptions and without acquiring an understanding of underlying principles (White and Horwitz, 1987).

Related work in Sweden (Marton, 1981; Entwistle and Marton, 1984; Dahlgren, 1984; Saljo, 1984) has used clinical interviews to explore the different understandings that students have of key principles and phenomena in a number of fields of learning. These interviews have revealed a range of student conceptions of each of the phenomena that the studies have explored, and have illustrated the importance of forms of learning which produce "a qualitative change in a person's conception of a phenomenon" from a lower-level, more naive conception to a more expert understanding of that phenomenon (Johansson, Marton, & Svensson, 1985, 235).

Under this view of learning, a student is rarely considered to have no understanding or no strategy when addressing a problem. Even beginning learners are considered to be engaged in an active search for meaning, constructing and using naive representations or models of subject matter. Rather than being "wrong", these representations frequently display partial understanding and are applied rationally and consistently by the individuals who use them. In arithmetic, for example, "it has been demonstrated repeatedly that novices who make mistakes do not make them at random, but rather operate in terms of meaning systems that they hold at a given time" (Nesher, 1986; also see Brown & Burton, 1978).

An implication of this view of learning for the assessment and monitoring of student learning is that much greater cognizance must be taken of the understandings and models that individual students construct for themselves during the learning process. In many areas of learning, levels of achievement might be better defined and measured not in terms of the number of facts and procedures that a student can reproduce, but in terms of his or her levels of understanding of the key concepts and principles that underlie a learning area (Glaser, 1981; Glaser, Lesgold, & Lajoie, 1987; Greeno, 1976).

An example of a study that has investigated students' levels of understanding is Carpenter and Moser's (1984) study of children's arithmetic skills. Carpenter and Moser found that most children in the first to third grades of school are able to provide correct answers to single-digit addition questions like 6+8=?. But children have a variety of methods of answering questions of this kind (see Table 1). These different methods indicate different levels of understanding and proficiency in single-digit addition. Some children solve 6+8=? by counting out six objects and another eight objects, and then counting all 14 (category 1). Later, children reach an understanding that counting does not have to begin at the number one. They "count on", although not necessarily from the larger number (e.g., "6; 7,8,...,14"; category 2). Later still, children understand the commutative property of addition (6+8 = 8+6) and consistently count on from the larger number ("8; 9,10,...,14"; category 3). Finally, by third grade, many children can solve 6+8=? using number facts, without counting objects (category 4). To monitor developing competence in single-digit addition, it is not sufficient to record only whether or not a child can provide the correct answer to a question like 6+8=?. By keeping track of the strategy that a child uses it is possible to infer the kinds of understanding that she or he has developed.

Insert Table 1 about here

This paper considers the problem of constructing measures of achievement that are based not on tests of learners' abilities to recall facts and apply memorized routines, but on inferences about students' levels of understanding of key concepts in an area of learning. Particular attention is given to the requirements of an achievement testing methodology if it is to be consistent with a view of learning as an active, constructive process.

2. CONVENTIONAL ACHIEVEMENT TESTING

Techniques for constructing achievement tests have been developed and refined over many decades. Most achievement tests begin with a statement of the instructional objectives to be assessed by each test. According to Bloom, Hastings, and Madaus (1971, 28), these objectives should be stated as directly observable student behaviors which can be reliably recorded as either present or absent. They should be "stated in terms which are operational, involving reliable observation and allowing no leeway in interpretation". To achieve this degree of reliability, test constructors are encouraged to write items to assess students' abilities to perform unambiguous, observable tasks such as "stating," "listing," "naming," "selecting," "recognizing," "matching," and "calculating" (Bloom et al., 1971, 34).

This emphasis on specifying and testing precise student behaviors has led to the construction of achievement tests composed of discrete items, each relating to a particular behavioral objective. and each scorable as either right or wrong. Multiple choice items have become especially popular in achievement tests because they can be scored quickly, unambiguously, and even by machine. In some areas of education, machine-scored multiple choice tests have become the principal mode of evaluating student learning. A disadvantage of conventional achievement tests is that, through their emphasis on precisely-defined student behaviors, they can encourage students to focus their efforts on relatively superficial forms of learning (Frederiksen, 1984).

In parallel with these developments in the practice of educational measurement, psychometric methods have been developed for the analysis of students' performances on test items of this kind. These methods have been introduced to transform records of right and wrong answers into measures of achievement, and to evaluate the reliability and validity of these measures. The more complex analytical methods, based on item response theory (IRT), take into account not only differences in the difficulties of individual test items, but also differences in their discriminating powers and, in the case of multiple choice

items, differences in their probabilities of being guessed correctly (Lord, 1980). Under IRT as well as under classical test theory, however, examinees' scores are essentially summaries of their tendencies to make correct rather than incorrect answers.

The alternative to conventional achievement testing discussed in this paper begins with a consideration of the key concepts, principles and phenomena that underlie a course of instruction and around which factual learning can be organized. Rather than recording students' understandings of these concepts as simply "right" or "wrong", this alternative approach recognizes that learners have a variety of understandings of phenomena, and that some of these understandings are less complete than others. The purpose of assessment is not to establish the presence or absence of specific behaviors, but to infer the nature of students' understandings of particular phenomena. Consequently, systems of observation very different from collections of distinct and conceptually isolated multiple choice test items are required.

3. BUILDING ACHIEVEMENT TESTS AROUND KEY CONCEPTS

The construction of an achievement test usually begins with a table of specifications with subject matter on one axis, and types of learning outcomes on the other. Items are written to cover outcomes like "knowledge of terminology," "knowledge of specific facts," and "principles and generalizations." In the use of such a table, these outcomes are treated as different but equivalent: the aim is to write items to cover each. However, because of the requirement that items be based on observable behaviors that can be scored right or wrong, and because it is easier to write items to assess students' knowledge of facts and procedures than to assess their understandings of principles and generalizations, achievement tests tend to be tests of students' abilities to recall and apply factual knowledge.

The method being proposed here begins by identifying key concepts in an area of instruction and building assessment procedures around these. These are fundamental principles, understandings, and approaches that a course of instruction aims to develop.

The difference between this approach and the conventional practice of treating "knowledge of principles" as an instructional objective of much the same status as "knowledge of facts" or "knowledge of terms" is that this approach makes the assessment of conceptual understanding the primary focus of the testing procedure.

A second fundamental difference between this approach and the usual approach to achievement testing is the emphasis placed on understanding how students view and think about key concepts. Rather than comparing students' responses with a "correct" answer, the emphasis is on inferring the nature or level of understanding reflected in each student's response.

One area in which a great deal of work has been done to understand how students think about and approach phenomena in that of physics education. Studies in several countries have explored students' understandings of such concepts as acceleration (Trowbridge and McDermott, 1981), electric charge, enthalpy and entropy, force and motion (Viennot, 1979), gravitation (Champagne, Klopfer and Anderson, 1980; Gunstone and White, 1981), light and the transmission of heat, momentum, potential difference, proportionality, torque, and such principles and models as Newton's laws, conservation laws, the atomic model, and electron flow models for circuits.

A common technique in these studies has been to ask students to describe what is happening in drawings of simple physical systems (e.g., to predict what will happen to an object, to describe the forces acting on a body, or to draw the trajectory that an object will follow). During these interviews, students are asked to explain their responses and their explanations are tape recorded (Johansson, Marton, & Svensson, 1985; McCloskey, 1983). In other studies, students have been asked to manipulate an apparatus in a laboratory to achieve particular effects (e.g., to apply a force to make a body move in a particular direction), while their explanations and comments are tape recorded and later transcribed (McDermott, 1984). Still other researchers (e.g., diSessa, 1982; White, 1983) have developed interactive software for this purpose. In these studies, students are asked

to apply "forces" to simulated objects on a screen to make them move to specified positions, to speed up, to slow down, and so on.

An example of the kind of question posed in these studies, taken from the work of McDermott (1984), is shown in Figure 1. In this study, students were presented with a drawing of a pendulum and asked to draw the trajectory that the weight would follow if the string of the pendulum broke when it was midway through its swing (i.e, in the vertical position). Four of the trajectories commonly drawn by students are shown in Figure 1.

Insert Figure 1 about here

Drawings B, C and D are all incorrect, but they reflect different levels of understanding. Drawings B and C show some understanding that the object will continue moving to the right after the string breaks (Newton's first law). Students who draw trajectory D show no understanding of this and recognize gravity as the only influence on the object's trajectory. Drawing B is almost correct: these students do not understand that the combination of a constant horizontal velocity and a vertical acceleration will be a parabolic trajectory. Drawing C shows the object continuing in the upward path that it would have followed had the string not been cut, and then falling under the influence of gravity. This drawing suggests a naive "impetus" theory of motion, a commonly held belief that an object will continue in its path (even a curved path) after the removal of the force that kept it moving in that path, until the object's "impetus" dissipates.

The observations made in these studies suggest that students do not simply make "random errors" but operate in terms of naive theories about physical phenomena. In the area of force and motion, these theories can be "remarkably well-articulated, ... quite consistent across individuals, ... and strikingly inconsistent with the fundamental principles of classical mechanics" (McCloskey, 1983, 299). In his studies of students' attempts to control a simulated object on a screen, diSessa (1982, 38) found "a surprising structure of

discrete and definite theories" about how forces influence motion. And, through their interviews with Swedish students about aspects of science learning, Johansson et al. (1985) arrive at a similar conclusion:

> In our case, a discovery of decisive importance was that for each phenomenon, principle. or aspect of reality, the understanding of which we studied. there seemed to exist a limited number of qualitatively different conceptions of that phenomenon, principle, or aspect of reality. (pp. 235-6)

A number of researchers have observed that the same naive conceptions can be found among students of different ages and with different educational backgrounds. McCloskey (1983), for example, found the same types of naive physical theories among students who had never taken physics, high school physics students, and college physics students. The only difference was in the frequencies of occurrence of these different understandings. McDermott (1984) reports an identical observation in a Norwegian study of high school physics students, future high school science teachers, and physics graduates.

A significant finding of these studies is that some students can succeed on traditional achievement tests and graduate from high school and even college physics courses with their naive conceptions of physical principles largely unchanged. Through their physics courses students are able to "master certain methods of calculation without having adopted the conceptualization underlying them" (Johansson et al., 1985, 235). Indeed, a misconception "may go undetected because a student's superficial knowledge of formulas and formula manipulation techniques can mask his or her misunderstanding of an underlying concept" (Clement, 1982, 66). The result is that "many students emerge from their study of physics and physical science without a functional understanding of some elementary but fundamental concepts" (McDermott, 1984, 31).

These findings invite a reconsideration of the way in which we think about and attempt to measure science learning. Clearly, many students are succeeding on precise,

operationally-defined objectives without developing an understanding of the material that they are learning. For many science educators, the answer is to place greater emphasis not on the learning of scientific facts and formulae, but on changing students' ways of thinking about scientific phenomena:

> The formal learning of science can be viewed as involving, at least in part, a shift from one set of beliefs about the physical world to another, one set of conceptions to another. (Osborne and Wittrock, 1985, 81).

and

> In our view, learning (or the kind of learning we are primarily interested in) is a qualitative change in a person's conception of a certain phenomenon or of a certain aspect of reality. (Johansson et al., 1985, 235).

4. CONSTRUCTING ORDERED OUTCOME CATEGORIES

Having identified key concepts in an area of learning and devised contexts (items) through which students' understandings of these concepts can be investigated, the next task is to delineate a set of categories for each item, through which student's observed responses are related to unobservable states of understanding. In this section and the two following, we address applications in which the most prevalent states of understanding can be ordered. This notion of order is basic to a view of learning as a "shift" in a student's understanding, with a shift constituting the desired "learning" when the change is from a lower level, more naive understanding to a higher level, more expert conception of a phenomenon.

This is not to say that all conceptions that students might bring to an item can be ordered from best to worst. We return later in the paper to consider some ways to model conceptions that differ but are not obviously more or less sophisticated. We begin here, however, by assuming the existence of a set of ordered categories for any given item (as will be illustrated below). For some items this set of categories might be constructed by grouping similarly sophisticated understandings. These constructed categories provide a

conceptual framework for recording an individual's response, and introduce the possibility of basing measures of achievement on inferences about students' levels of understanding.

Grouping students' responses to construct a set of categories of understanding is part of the method used by Marton (1981) and his colleagues at the University of Gothenburg. These researchers interview students to explore their understandings of particular concepts and principles, transcribe tape recordings of these interviews, and then carry out detailed analyses of transcripts. "The aim of the analysis is to yield descriptive categories representing qualitatively distinct conceptions of a phenomenon". These categories form an "outcome space" which provides "a kind of analytic map" of students' understandings of each phenomenon. Learning is thought of as "a shift from one conception to another" on this map (Dahlgren, 1984, 24-31).

Carpenter and Moser (1984) provide a picture of such a map. From their analysis of students' performances on single-digit addition tasks, they . constructed the five ordered outcome categories shown in Table 1. Children in category 0 are unable to solve 6+8=?. Children in category 1 understand that 6+8=? can be solved by counting the total number of objects in two groups of sizes 6 and 8. Children in category 2 also understand that the counting of objects does not have to begin at the number one, and so "count on." Children in category 3 understand the commutative property and count on from the larger number. Children in category 4 have a level of understanding that enables them to use number facts to solve 6+8=? without counting.

Figure 2 shows the proportion of a group of Wisconsin children in each of the five outcome categories at each of eight time points during their first three years of school. At the beginning of first grade (bottom of the map), about 15 percent of these children could not solve problems like 6+8=?, even with blocks (Category 0). Among those children who could solve such a problem, by far the most common strategy was to count out six objects and another eight objects and then to count all fourteen (Category 1). By the beginning of the second grade, almost all these children understood that counting does not have to begin

at the number one and were counting on (Categories 2 and 3), although some still did not understand the commutative property and were not counting consistently from the larger number. By the eighth round of observations (top of the map), more than 70 percent of this group of children could solve single-digit addition problems without having to count objects. Carpenter and Moser provide similar outcome maps for other aspects of addition and subtraction learning.

Insert Table 1 and Figure 2 about here

---

5. COLLECTING OBSERVATIONS

While conversations with students are probably essential for identifying the variety of understandings that learners have of phenomena and for constructing sets of outcome categories, interviews are not practicable as a basis for achievement testing. Alternative observation methods must be found which will permit inferences to be made about students' levels of understanding. These procedures must go deeper than identifying incorrect answers: they must attempt to identify the nature of the understandings and models that individual students are employing. In general, this will require imaginative new approaches to achievement testing.

One possible approach is the "rule assessment" procedure developed by Siegler (1978, 1981). This approach uses a carefully constructed set of questions designed to expose different levels of understanding of a concept. While each individual question might be scored as right or wrong, neither the response to any one item nor total score on a set of items are sufficient to differentiate students using different rules. Rather, it is a student's pattern of right and wrong answers that constitutes a basis for inferring his or her level of understanding.

Another approach is to use computer-administered tasks as the testing medium. This approach introduces the possibility of matching each student's response to a library of common responses rather than to a single "correct" answer. In the pendulum task in Figure 1, for example, students might be asked to draw a trajectory on a screen and each student's drawing might then be referred to a library of common student responses. In this way, a student's response might automatically be assigned to one of several ordered outcome categories for that task, and a record made of the student's apparent conception or theory concerning that phenomenon.

A decision about a student's assignment to an outcome category might be based on the students' responses to several related questions, looking for, in Brown and Burton's (1978) terminology, consistent "bugs" in their solutions. The automatic generation of hypotheses about students' understandings might be followed by further questions aimed at confirming those hypotheses. Does a student who draws trajectory C in Figure 1 also believe that an object fired out of a curved tube will continue in a curved path for a short time after leaving the tube? Through carefully designed hints and subquestions it may be possible to emulate in a crude way the type of exploration that can be done through an interview to trace a student's misunderstanding to its source. Ordered outcome categories, for example, might then be defined in terms of responses to a set of related questions or tasks.

In an achievement test of this type, tasks may bear little resemblance to traditional achievement test questions. As diSessa (1982) and White (1983) show, a great deal of information can be collected about individuals' naive theories of force and motion by asking them to move simulated objects on a screen. A computer can be used to keep detailed records of when students apply forces, in which directions they apply those forces, and how they respond to the motion that they produce. Automatic analyses of student records might be used to infer students' levels of understanding. Simulations of this kind could be used in a wide variety of learning areas—for example, the use of simulated patient

management problems to explore students' levels of understanding of medical principles and to expose inappropriate or potentially misleading ways of thinking about particular phenomena (of course, the analysis of these data would be far more complex than the simple examples given here).

6. CONSTRUCTING MEASURES OF ACHIEVEMENT

If the types of observations that result from these testing procedures are to provide a basis for achievement measurement and are to be a viable alternative to conventional achievement tests, then models and methods analogous to those that have been developed for right/wrong test questions are required to supervise the construction of the new measures.

The starting point in the development of a method for ordered outcome categories is a matrix of observations like the matrix shown in Table 2. This hypothetical data matrix shows the responses of 32 students to 8 items (e.g., Carpenter & Moser's single-digit addition items). Responses to each item are recorded in one of five ordered categories (labelled 0 to 4). Students' scores on each item have been arranged in this matrix in an orderly way with abrupt transitions between adjacent categories. (This can be seen by reading down each column.) The consequence of ordering scores on each item in this way is that it is possible to infer from the full data matrix in Table 2 an unambiguous order for these 32 students on the single achievement dimension defined by these eight items.

Insert Table 2 about here

It is unlikely that a perfectly orderly pattern of scores on an item will occur in practice. The transition from category x-1 to category x of an item is not likely to be sharp, as depicted in Table 2, but to be gradual. Rather than expecting a person above a particular level of ability in an area of learning to definitely score x rather than x-1 on an item, it is

more realistic to imagine a score of x becoming more likely than a score of x-1 at higher levels of ability. In other words, a probabilistic formulation will in general be more appropriate than a deterministic representation (see Wilson, 1989a).

The psychometric method described here, the Partial Credit Model (PCM; Masters, 1982; Wright and Masters, 1982), proposes that the probability of a person scoring x rather than x-1 on a particular item i will increase steadily with ability in an area of learning such that

$$\frac{\pi_{nix}}{\pi_{nix-1}+\pi_{nix}} = \frac{\exp(\theta_n-\delta_{ix})}{1+\exp(\theta_n-\delta_{ix})}, \qquad [1]$$

where π_{nix} is the probability of person n responding in category x (x=1,2,...,m_i) of item i, θ_n is person n's level of proficiency in the area of learning measured by this set of items, and δ_{ix} is a parameter associated with the transition between outcome categories x-1 and x of item i.

The consequence of applying the simple logistic expression [1] to the transition between each pair of adjacent outcome categories for each item, is that a connection is formed between the ordered categories for that item and the underlying variable that the set of items is used to measure. It is this connection that enables performances on each item to be used to estimate students' locations on the underlying variable. The nature of this probabilistic connection is illustrated in Figure 3, in terms of response probabilities for a hypothetical single digit addition problem.

Insert Figure 3 about here

Figure 3 shows how, under the PCM, the probability of a response in each category of an item changes with increasing student proficiency. It has been drawn to

resemble Figure 2. The difference is that Figure 3 does not show observed proportions of students in each category, but modelled proportions. For any given level of θ, one looks across the graph to determine the probabilities of a response in category at this level of proficiency. The basic shapes of the five zones in Figure 3 are fixed by the PCM and are the consequence of using the simple logistic expression [1] to model the transition between adjacent categories of each item. The widths and locations of the zones for each item are estimated from students' responses to that item, and are expressed through the δ parameters.

The probabilistic partial credit model depicted in Figure 3 enables measures of achievement to be constructed from inferences of students' levels of understanding of each of a number of concepts or phenomena in an area of learning. A student's θ parameter indicates not simply a tendency to make correct responses, but tendencies to provide answers reflecting the various levels of understanding on a collection of tasks probing that understanding. The model serves the same function in the analysis of responses recorded in ordered outcome categories as the item response models that have been developed for dichotomously-scored responses (Rasch, 1960; Lord and Novick, 1968; Lord, 1980), summarizing, in terms of the task and person parameters, the patterns in the data that are consonant with a conception of student proficiency. Estimation procedures and tests of model-data fit for the PCM are described by Wright and Masters (1982). Tests of item fit (which can be thought of as comparisons of the observed and modelled maps for an item as shown in Figures 2 and 3) provide internal consistency indices analogous to traditional item statistics like biserial correlations. Tests of person-fit flag occurrences of unusual response patterns, as might occur when a student's state of understanding is atypical, and requires special attention.

7. PARTIALLY-ORDERED STATES

The psychometric model just described can be used when a set of ordered categories is defined for each item. However, attempting to order all conceptions of a phenomenon from "worst" to "best" may not always be fruitful. In some cases, two or more ways of visualizing a problem can be identified, none better or worse than another. If these different conceptions have different implications for instruction, than maintaining a distinction among them can be useful.

Norman's (1983) and Gentner and Gentner's (1983) studies of students' models for electrical circuits provides an example. These studies suggests that many students visualize electric circuits in terms of more familiar physical systems. Some, for example, see electric current as analogous to water flow. Batteries are visualized as reservoirs, and resistors correspond to constrictions in water flow. This analogy facilitates the solution of problems about power sources in parallel and series, but impedes solutions to problems about parallel and series resistors. Other students see an electrical power source as analogous to a crowd entering a stadium, with resistors as turnstiles through which they must pass. This "teeming crowd" analogy facilitates problems about combinations of resistors, but offers little insight into battery combinations.

Each of these models captures some aspects of electrical systems. Students using either model have a better understanding than students with no model at all. On the other hand, neither of these physical models provides a complete understanding of current flow or of the operation of circuits. A higher level of understanding requires an appreciation of the limitations of the physical analogies as models for circuits. In this sense, students who operate with either one of the two models can be thought of as being at similarly intermediate levels of understanding.

From the point of view of traditional test theory and the maximization of test reliability, it is difficult to justify distinguishing among students who use the water flow analogy and those who use the teeming crowds analogy. Items that distinguish between

these two groups are likely to contribute little to reliability, as their discriminating power is among people at similar levels of overall proficiency. But further instruction might well differ for the two groups—first explicating the model that a student's responses suggest he or she may be using (perhaps intuitively), exploring its uses and limitations, then introducing the complementary model and its sphere of usefulness.

To develop a model for these situations, let us suppose that we can identify K states of understanding in a learning area, subsets of which may be ordered, but others of which may not be. Items are characterized by identifiable features that determine their difficulties within these states. In the electrical circuits example, for instance, resistor problems are relatively easier than battery problems for students using the teeming crowds analogy, while the battery problems are relatively easier for those using the water flow analogy. From each student's responses, we wish to infer his or her state of understanding (ϕ_n, which ranges from 1 to K) and degree of proficiency within that state (θ_n).

The essence of this approach is that while a single proficiency summary of performance fails to characterize important differences among learners, it may suffice in some applications to use a single proficiency to characterize differences among learners in the same type of understanding, while further distinguishing among these qualitative states. The fact that these variables can never be known with certainty is reflected by the nature of the inferences that are drawn about students: probabilities that the student is in the possible states, and an estimate of proficiency corresponding to each possibility.

The details of such models are given by Mislevy and Verhelst (1990). In the case of items scored right or wrong, the probability of a correct response to Item i from Person n, who is in state k of understanding (ϕ_n=k) and has proficiency θ_n, is given as:

$$P(x_{ni}=1 | \theta_n, \phi_n=k, \beta_{ik}) = f_k(\theta_n,\beta_{ik}) \ , \qquad [2]$$

where β_{ik} characterizes such features of Item i as its difficulty and f_k is a function relating examinee and item parameters to probabilities of correct response—both as pertain to persons in level k only. When persons from only one level are under consideration, [2] is a standard IRT model. The item parameters β_{ik} can be expected to vary from one level of understanding to the next, however—and indeed they must vary if the model is to be practically useful for distinguishing students at one level from those at another.

To illustrate the approach, we present highlights of a one of many aspects of an analysis carried out by Wilson (1984), using Robert Siegler's (1978, 1981) data and rule-acquisition perspective. For additional examples, the reader is referred to Mislevy and Verhelst (1990), Mislevy, Wingersky, Irvine, and Dann (in press), and Wilson (1989b).

Figure 4 shows two of Siegler's six balance beam problem prototypes. In E ("Equal") items, both the weights and distances are the same on the two sides of the scale, and the correct answer is that the beam will balance. In S ("Subordinate") items, the same numbers of weights are on both sides, but on one side they are further from the fulcrum. That side will tip down. Following Piaget (Inhelder & Piaget, 1958; Piaget, 1960), Siegler posits that children typically exhibit distinct stages as they acquire competence in proportional reasoning, adding to their repertoire the increasingly sophisticated rules listed in Table 3. Children can thus differ as to their stage of understanding, or their proficiency in using the rules they currently command. In particular, a qualitative shift occurs when a child apprehends the salience of distance in balance beam problems. Before this realization, children see no systematic, relevant, differences between E and S items, and tend to predict the beam will balance in both situations.

Insert Table 3 and Figure 4 about here

Among other analyses, Wilson (1984) analyzed responses to four E and four S items from two perspectives. The first was based on the Rasch IRT model for right/wrong

items. Under the Rasch model, the probability that Person n will respond correctly to Item i is a function of the person's proficiency parameter, θ_n, and the item's difficulty parameter, β_i:

$$P(x_{ni}=1|\theta_n,\beta_i) = \frac{\exp(\theta_n-\beta_i)}{1+\exp(\theta_n-\beta_i)} . \qquad [3]$$

(Note the similarity of [3] to [1]; the Rasch model for right/wrong items is a special case of the PCM). Figure 5 illustrates the results. The relative positions of an item and a person on the scale (θ_n-β_i) determine the probability of a correct response through [3]. Not surprisingly, S items are seen to be harder than E items. If the Rasch model were correct, increasing competence would be reflected in similar increases in the chances of correct response to both E and S items. But analyses of person-fit to the Rasch model revealed relatively fewer correct answers to S items from many children who did well on E items, and relatively fewer incorrect answers to E items from children who did well on S items, than would be expected under the Rasch model.

Wilson resolved these anomalies in the second analysis, based on his "Saltus" (Latin for "leap") model for development that occurs in stages. Saltus extends the Rasch model by incorporating stage membership parameters for persons and "Saltus parameters" that allow for discontinuities such as the transition from Rule I to Rule II. In this analysis, children who <u>had not</u> experienced the transition were modeled in accordance with [3]; those who <u>had</u> were modeled by a model of the same form, but with the Saltus parameter τ subtracted from the difficulty parameters of S items. In terms of Equation [2], f_I and f_{II} both have the functional form given in [3], $\beta_{iII}=\beta_{iI}$ for E items, and $\beta_{iII}=\beta_{iI}-\tau$ for S items. Figure 6 illustrates the effect. In effect, τ measures the quantitative effect on performance associated with a qualitative change in understanding.

Insert Figures 5 and 6 about here

---

8. OTHER APPROACHES

The psychometric literature has begun to offer models that might be used to guide the construction and analysis of achievement tests of the kind proposed here. Some are mentioned below.

Wilson's (1984, 1989b) Saltus model for hierarchical stages of development (illustrated above) provides a stochastic framework for psychological models such as Piaget's (1960) and Siegler's (1978, 1981) that posit predictable discontinuities in proficiencies as development occurs, and educational models such as Gagné's (1968) and Riley's (Riley, 1981; Riley, Greeno, & Heller, 1983) that posit detectable patterns of task difficulties as students progress through successive levels of competence.

Latent class models (e.g., Haertel, 1984, 1989; Haertel& Wiley, in press; Macready and Dayton, 1980) accommodate nonordered states of competence and reconfigurations of proficiencies, without further differentiating students within a state. Computational limitations to less than about ten items per student have all but precluded their use for measuring individual achievement. Recent developments by Paulson (1985) and Yamamoto (1987) enable the use of these models with up to sixty items, opening the door to precise estimation for individual students and even potentially adaptive testing (Macready & Dayton, 1989).

Yamamoto (Yamamoto, 1987; Yamamoto & Gitomer, in press) has also introduced a "hybrid" model for a mixture of latent classes and an IRT class. No claim is made that such a mixture accurately reflects the psychological reality of students' behavior, but a practical advantage is emphasized: Explicit classes can be defined to correspond to available instructional options while an amorphous IRT class accounts for potentially large

numbers of remaining classes, distinctions among which are irrelevant to the decision that must be made.

Another approach that leans on IRT to handle bookkeeping tasks in complex problems is Kikumi Tatsuoka's (K.K. Tatsuoka,1983, 1989; K.K. Tatsuoka & M.M. Tatsuoka, 1987) "rule space" procedure. A standard IRT model is first fit to item responses. If the IRT model were correct, estimates of persons' proficiency would account for all systematic patterns within the data. But Tatsuoka then calculates an index of lack of fit from the IRT model, and studies the joint distribution of proficiency estimates under the IRT model and indices of lack of fit from that model. The ordered pairs of proficiency estimates and fit indices often suffice to identify systematic patterns of response that correspond to particular solution strategies, thereby identifying users of particular erroneous rules as well as correct rules.

Embretson's (1985, in press) model for multiple strategies requires identifying different sequences of component subtasks that can be used to solve problems. This approach can be applied when it is possible to observe the results of subtask operations as well as a global result, and, as such, is amenable to procedures described above which enable the definition of levels of understanding for identified composite tasks. If levels of understanding are ordered, the results of microanalyses using Embretson's model could serve as input to achievement measurement via the partial credit model.

Our discussions and examples have addressed relatively simple situations, with a single developing concept with just a few stages. As such, however, they constitute building blocks for characterizing students' knowledge with respect to larger systems of interconnected concepts. The interested reader is referred to Mislevy, Yamamoto, and Anacker (in press) on the possibility of constructing Bayesian inference networks for this purpose.

9. CONCLUSION

Recent developments in cognitive and educational psychology reveal that most meaningful learning contrasts markedly to the type of learning implied by standard psychometric procedures—those based on item response theory as well as those using classical true-score test theory. The difference is characterized by the discontinuities of real-world learning, as learners reconfigure their knowledge, combine existing skills in new ways, and develop alternative strategies for solving problems.

It is possible to build achievement tests that measure learning of this kind. It is not possible to do so with traditional item writing rules, test construction procedures, and scoring formulas. To operationalize the new approach, the structure of learning is integral at each step along the way, from writing items through reporting achievement. In return for this greater investment in the psychology of the learning area, one can expect a greater utility: a measure of achievement which, by reflecting the nature of competence as attained thus far, sets the stage for further learning.

REFERENCES

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). **Handbook on formative and summative evaluation of student learning.** New York: McGraw-Hill.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural errors in basic mathematical skills. **Cognitive Science, 2,** 155-192.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. **Cognition, 9,** 117-123.

Carpenter, T.P. & Moser, J.M. (1984). The acquisition of addition and subtraction concepts in grades one through three. **Journal for Research in Mathematics Education, 15,** 179-202.

Champagne, A.B., Klopfer, L.E. & Anderson, J.H. (1980). Factors influencing the learning of classical mechanics. **American Journal of Physics, 48,** 1074-1079.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. **Cognitive Psychology, 4,** 55-81.

Chi, M.T.H., Feltovich, P.J. & Glaser, R. (1981) Categorization and representation of physics problems by experts and novices. **Cognitive Science, 5,** 121-152.

Clement, J. (1982). Students' preconceptions of introductory mechanics. **American Journal of Physics, 50,** 66-71.

Dahlgren, L-O. (1984). Outcomes of learning. In Marton, F., Hounsell, D. & N. Entwistle (Eds.), **The experience of learning.** Edinburgh: Scottish Academic Press.

diSessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. **Cognitive Science, 5,** 37-75.

Driver, R, & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. **Studies in Science Education, 5,** 61-84.

Embretson, S.E. (1985). Multicomponent latent trait models for test design. In Se.E. Embretson (Ed.), **Test design: Developments in psychology and psychometrics.** Orlando: Academic Press.

Embretson, S.E. (in press). Psychometric models for learning and cognitive processes. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), **Test theory for a new generation of tests.** Hillsdale, NJ: Erlbaum.

Entwistle, N., & Marton, F. (1984). Changing conceptions of learning and research. In Marton, F., Hounsell, D. & N. Entwistle (Eds.), **The experience of learning.** Edinburgh: Scottish Academic Press.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. **American Psychologist, 39,** 193-202.

Gagné, R.M. (1968). Learning hierarchies. **Educational Psychologist, 6,** 1-9.

Gentner, D., & Gentner, D.R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. Stevens (Eds.), **Mental models.** Hillsdale, NJ: Erlbaum.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. **American Psychologist, 36,** 923-936.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), **The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing** (Vol. 3). Hillsdale, NJ: Erlbaum.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), **Cognition and instruction.** Hillsdale, NJ: Erlbaum.

Gunstone, R., & White, R. (1981). Understanding of gravity. **Science Education, 65,** 291-299.

Haertel, E.H. (1984). An application of latent class models to assessment data. **Applied Psychological Measurement, 8,** 333-346.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement test items. **Journal of Educational Measurement, 26,** 301-321.

Haertel, E.H., & Wiley, D.E. (in press). Poset and lattice representations of ability structure: Implications for test theory. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), **Test theory for a new generation of tests.** Hillsdale, NJ: Erlbaum.

Inhelder, B., & Piaget, J. (1958). **The growth of logical thinking from childhood to adolescence.** New York: Basic.

Johansson, B., Marton, F., & Svensson, L. (1985). An approach to describing learning as change between qualitatively different conceptions. In West, L.H. & L.A. Pines (Eds.), **Cognitive structure and conceptual change.** Orlando, Florida: Academic Press.

Lesgold, A.M., Feltovich, P.J., Glaser, R., & Wang, Y. (1981). The acquisition of perceptual diagnostic skill in radiology (Technical Report No. PDS-1). Pittsburgh: Learning Research and Development Center, University of Pittsburgh.

Lord, F.M. (1980). **Applications of item response theory to practical testing problems.** Hillsdale, NJ: Erlbaum.

Lord, F.M. & Novick M.R. (1968). **Statistical theories of mental test scores.** Reading, Massachusetts: Addison Wesley.

Macready, G.B., & Dayton, C.M. (1980). The nature and use of state mastery models. **Applied Psychological Measurement, 4,** 493-516.

Macready, G.B., & Dayton, C.M. (1989, March). Adaptive testing with latent class models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Marton, F. (1981). Phenomenography—describing conceptions of the world around us. **Instructional Science, 10,** 177-200.

Masters, G.N. (1982). A Rasch model for partial credit scoring. **Psychometrika, 47,** 149-174.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. Stevens (Eds.), **Mental models.** Hillsdale, NJ: Erlbaum.

McDermott, L.C. (1984). Research on conceptual understanding in mechanics. **Physics Today,** 1-10.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. **Psychometrika, 55,** 195-215.

Mislevy, R.J., Wingersky, M.S., Irvine, S.H., and Dann, P.L. (in press). Resolving mixtures of strategies in spatial visualization tasks. **British Journal of Mathematical and Statistical Psychology.**

Mislevy, R.J., Yamamoto, K, & Anacker, S. (in press). Toward a test theory for assessing student understanding. In R.A. Lesh (Ed.), **Assessing higher-level understanding in middle-school mathematics.** Hillsdale, NJ: Erlbaum.

Nesher, P. (1986) Learning mathematics: A cognitive perspective. **American Psychologist, 41.** 114-1122.

Norman, D.A. (1983). Some observations on mental models. In D. Gentner & A. Stevens (Eds.), **Mental models.** Hillsdale, NJ: Erlbaum.

Osborne, R.J. & Gilbert, J.K. (1980). A technique for exploring students' views of the world. **Physics Education, 15,** 376-379.

Osborne, R.J., & Wittrock, M.C. (1983). Learning science: A generative process. **Science Education, 67,** 489-508.

Osborne, R.J., & Wittrock, M.C. (1985). The generative learning model and its implications for science education. **Studies in Science Education, 12,** 59-87.

Paulson, J. (1985). Latent class representations of systematic patterns in test responses. ONR Technical Report. Portland: Portland State University.

Piaget, J. (1960) The general problems of the psychological development of the child. In J.M. Tanner and B. Inhelder (Eds.), **Discussions on Child Development: Vol. 4. The fourth meeting of the World Health Organization Study Group on the Psychological Development of the Child, Geneva, 1956.**

Posner, G.J., Strike, K.A., Hewson, P.W., & Gertzog, W.A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. **Science Education, 66,** 211-227.

Rasch, G. (1960). **Probabilistic models for some intelligence and attainment tests.** Copenhagen: Danish Institute for Educational Research.

Riley, M.S. (1981). Conceptual and procedural knowledge in development. Unpublished Master's thesis, University of Pittsburgh.

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H.P. Ginsburg (Ed.), **The development of mathematical thinking** (pp. 153-196). New York: Academic Press.

Saljo, R. (1984). Learning from reading. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), **The experience of learning.** Edinburgh: Scottish Academic Press.

Siegler, R.S. (1978). The origins of scientific reasoning. In R.S. Siegler (Ed.), **Children's Thinking: What Develops?** Hillsdale, N.J.: Erlbaum.

Siegler, R.S. (1981). Developmental sequences within and between concepts. **Monograph of the Society for Research in Child Development, 46.**

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. **Journal of Educational Measurement, 20,** 345-354.

Tatsuoka, K.K. (1989). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), **Diagnostic monitoring of skill and knowledge acquisition.** Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K., & Tatsuoka, M.M. (1987). Bug distribution and statistical pattern classification. **Psychometrika, 52**, 193-206.

Trowbridge, D.E., & McDermott, L.C. (1981). Investigation of student understanding of the concept of acceleration in one dimension. **American Journal of Physics, 49**, 242-253.

Viennot, L. (1979). Spontaneous reasoning in elementary dynamics. **European Journal of Science Education, 1**, 205-221.

Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem-solving skill in the social sciences. In G.H. Bower (Ed.), **The psychology of learning and motivation: Advances in research and theory** (Volume 17, pp. 165-213). New York: Academic Press.

White, B.Y. (1983). Sources of difficulty in understanding Newtonian dynamics. **Cognitive Science, 7**, 41-65.

White, B.Y. & Horwitz, P. (1987) ThinkerTools: Enabling children to understand physical laws. **Proceedings of the Ninth Annual Conference of the Cognitive Science Society.** Hillsdale, NJ: Erlbaum.

Wilson, M.R. (1984) **A Psychometric Model of Hierarchical Development.** Doctoral dissertation, University of Chicago.

Wilson, M.R. (1989a). A comparison of deterministic and probabilistic approaches to measuring learning structures. **Australian Journal of Education, 33**, 125-138.

Wilson, M.R. (1989b). Saltus: A psychometric model of discontinuity in cognitive development. **Psychological Bulletin, 105**, 276-289.

Wright, B.D. & Masters, G.N. (1982). **Rating scale analysis.** Chicago: MESA.

Yamamoto, K. (1987) **A hybrid model for item responses.** Doctoral dissertation, University of Illinois.

Yamamoto, K., & Gitomer, D. H. (in press). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), **Test theory for a new generation of tests.** Hillsdale, NJ: Erlbaum.

Table 1

Outcome Categories for Single-Digit Addition

(e.g., 6+8 = ?)

Category	Description
4	Does not need to count objects, but uses number facts to solve 6+8 = 14.
3	Always counts.on from the larger number ("8; 9,10,...,14").
2	Counts on, but not consistently from the larger number ("6; 7,8,...,14").
1	Counts out 6 objects and 8 objects and then counts them all ("1.2,...,14").
0	Unable to solve.

Table 2

Hypothetical Data Matrix for Single-digit Addition

	Items							
Students	1	2	3	4	5	6	7	8
1	4	4	4	4	4	4	4	4
2	4	4	4	4	4	4	3	4
3	4	4	4	4	4	3	3	3
4	4	4	4	4	4	3	3	3
5	4	4	4	4	4	3	3	2
6	4	4	4	3	4	3	3	2
7	4	4	4	3	4	3	2	2
8	4	4	4	3	3	3	2	2
9	4	4	4	3	3	3	2	1
10	4	4	4	3	3	2	2	1
11	4	4	3	3	3	2	2	1
12	4	3	3	3	3	2	2	1
13	4	3	3	2	3	2	2	1
14	4	3	3	2	3	2	2	0
15	3	3	3	2	3	2	2	0
16	3	3	3	2	3	2	1	0
17	3	3	3	2	2	2	1	0
18	3	3	2	2	2	2	1	0
19	2	3	2	2	2	2	1	0
20	2	3	2	1	2	2	1	0
21	2	2	2	1	2	2	1	0
22	2	2	2	1	2	1	1	0
23	2	2	1	1	2	1	1	0
24	2	2	1	1	1	1	1	0
25	2	2	1	1	1	1	0	0
26	2	2	1	1	0	1	0	0
27	2	2	1	1	0	0	0	0
28	2	1	1	1	0	0	0	0
29	1	1	1	1	0	0	0	0
30	1	1	0	0	0	0	0	0
31	1	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0

Note: Table entries are observed outcome categories, coded from 0 to 4.

Table 3

Hierarchy of Rule Acquisition

Rule	Description
Rule 0	Salience of neither weight nor distance recognized; answers depend on personal factors.
Rule I	If the weights on both sides are equal, it will balance. If they are not equal, the side with the heavier weight will go down. (Weight is the "dominant dimension," because children are generally aware that weight is important in the problem earlier than they realize that distance from the fulcrum, the "subordinate dimension," also matters.)
Rule II	If the weights and distances on both sides are equal, then the beam will balance. If the weights are equal but the distances are not, the side with the longer distance will go down. Otherwise, the side with the heavier weight will go down. (A child using this rule uses the subordinate dimension only when information from the dominant dimension is equivocal.)
Rule III	Same as Rule II, except that if the values of both weight and length are unequal on both sides, the child will "muddle through" (Siegler, 1981, p.6). (A child using this rule now knows that both dimensions matter, but doesn't know just how they combine. Responses may be based on a strategy such as guessing.)
Rule IV	Combine weights and lengths correctly (i.e., compare torques, or products of weights and distances).

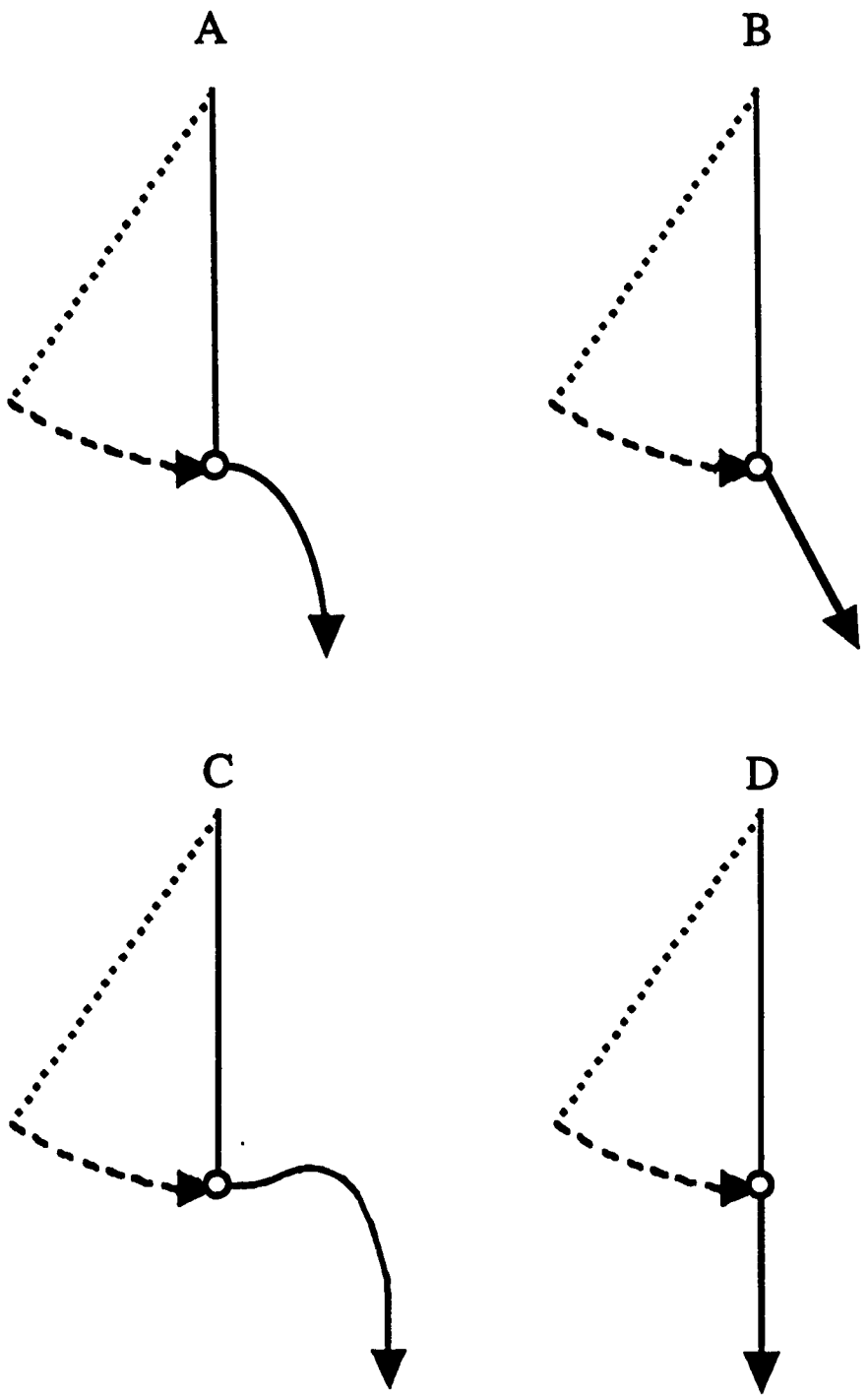


Figure 1

Common Responses to a Physics Task

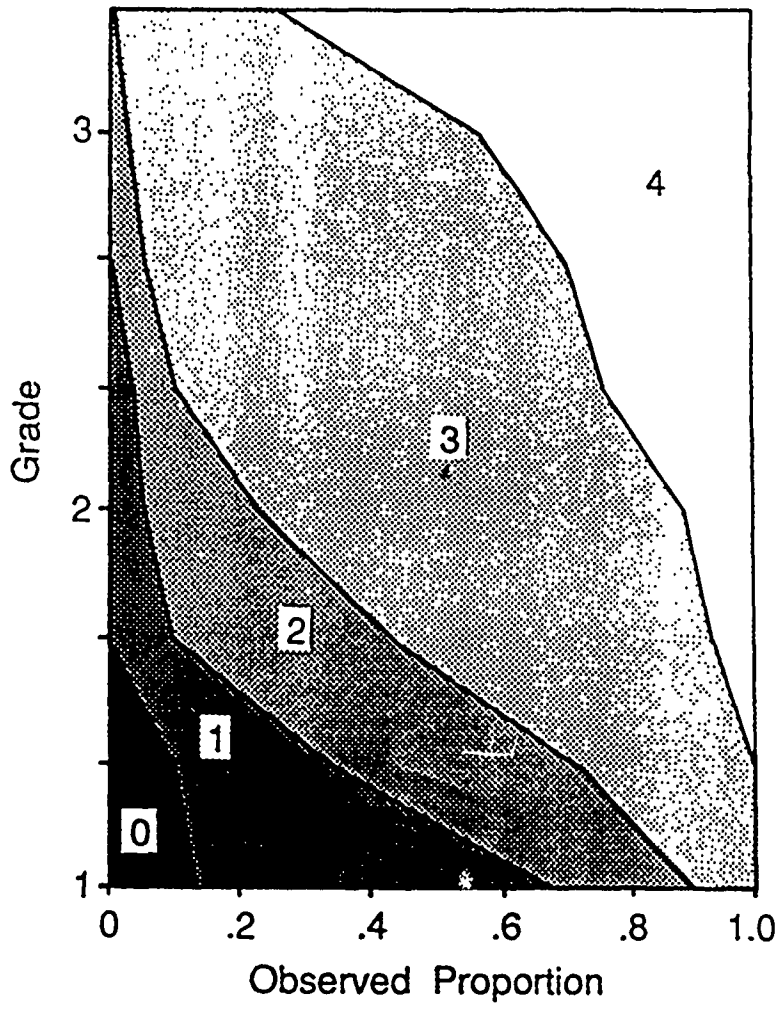


Figure 2

Observed Proportions of Children in Each of Five Ordered Outcome Categories on a Single-digit Addition Item

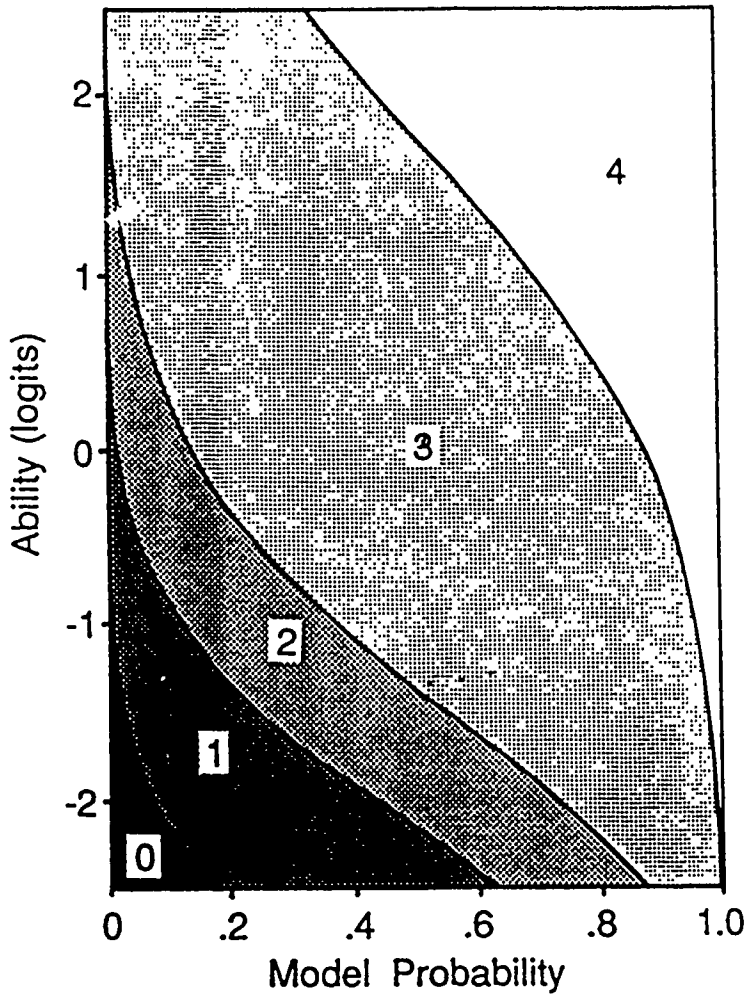


Figure 3

Modeled Probabilities of Responding in Each of Five Ordered Outcome Categories on a Single-digit Addition Item

Will the beam tip left, tip right, or stay flat
when the gray blocks are taken away?

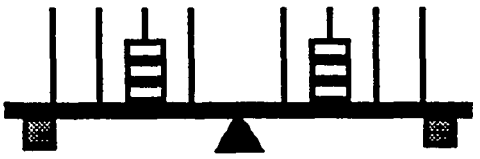
Item Type	Sample Item
E	
S	

Figure 4
Prototypical Balance Beam Items

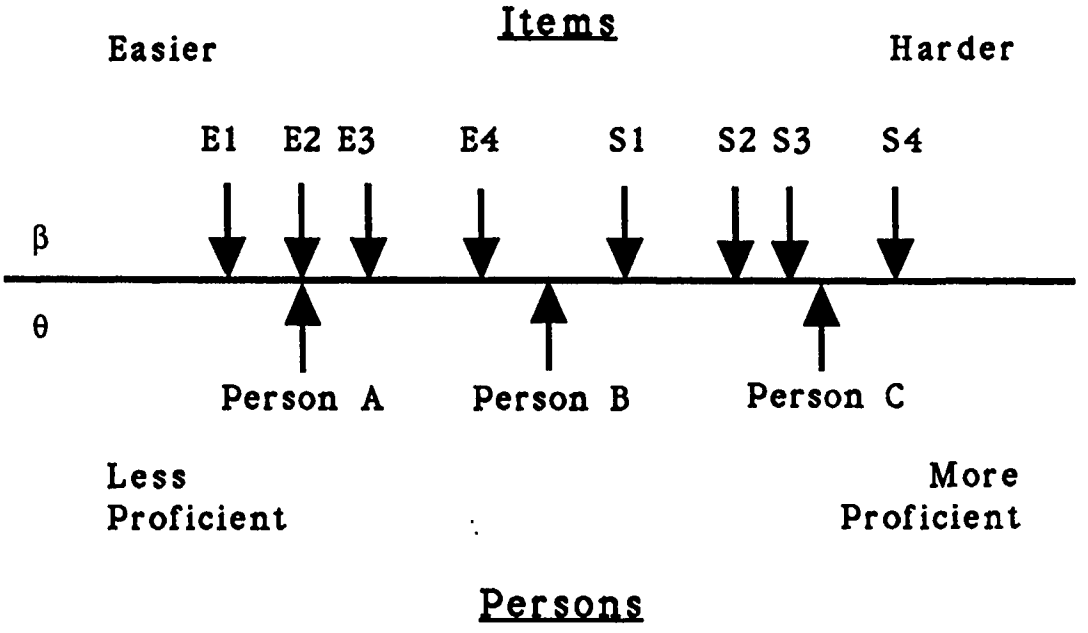
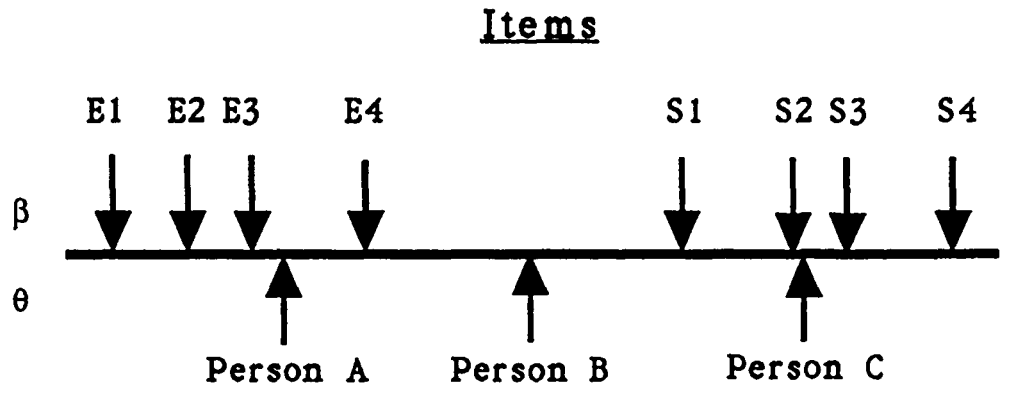


Figure 5

Rasch Model Representation
of Balance Beam Items



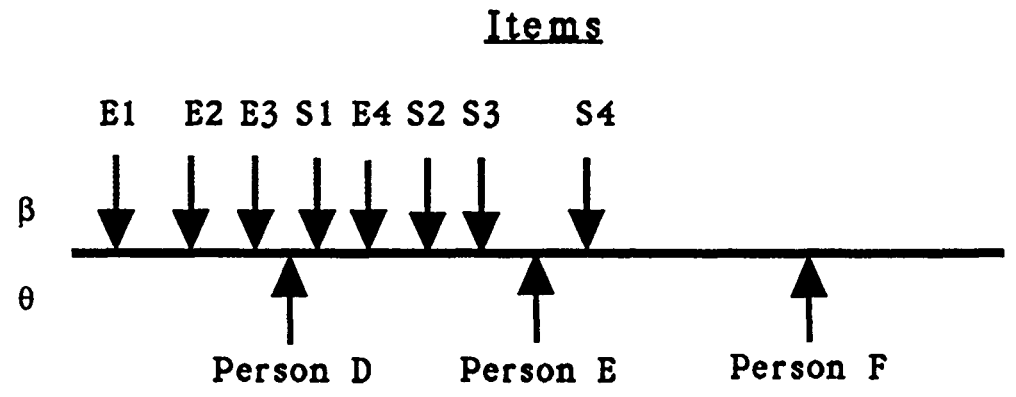


Figure 6

Saltus Model Representation
of Balance Beam Items

Dr. Terry Ackerman
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. James Algina
1403 Norman Hall
University of Florida
Gainesville, FL 32605

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Dr. Ronald Armstrong
Rutgers University
Graduate School of Management
Newark, NJ 07102

Dr. Eva L. Baker
UCLA Center for the Study
of Evaluation
145 Moore Hall
University of California
Los Angeles, CA 90024

Dr. Laura L. Barnes
College of Education
University of Toledo
2801 W. Bancroft Street
Toledo, OH 43606

Dr. William M. Bart
University of Minnesota
Dept. of Educ. Psychology
330 Burton Hall
178 Pillsbury Dr., S.E.
Minneapolis, MN 55455

Dr. Isaac Bejar
Mail Stop: 10-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Ramat Aviv 69978
ISRAEL

Dr. Arthur S. Blaiwes
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekruterings-En Selectiecentrum
Kwartier Koningen Astrid
Bruijnstraat
1120 Brussels, BELGIUM

Dr. Robert Breaux
Code 281
Naval Training Systems Center
Orlando, FL 32826-3224

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. Gregory Candell
CTB/McGraw-Hill
2500 Garden Road
Monterey, CA 93940

Dr. John B. Carroll
409 Elliott Rd., North
Chapel Hill, NC 27514

Dr. John M. Carroll
IBM Watson Research Center
User Interface Institute
P.O. Box 704
Yorktown Heights, NY 10598

Dr. Robert M. Carroll
Chief of Naval Operations
OP-01B2
Washington, DC 20350

Dr. Raymond E. Christal
UES LAMP Science Advisor
AFHRL/MOEL
Brooks AFB, TX 78235

Mr. Hua Hua Chung
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director, Manpower Program
Center for Naval Analyses
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Director,
Manpower Support and
Readiness Program
Center for Naval Analyses
2000 North Beauregard Street
Alexandria, VA 22311

Dr. Stanley Collyer
Office of Naval Technology
Code 222
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Hans F. Crombag
Faculty of Law
University of Limburg
P.O. Box 616
Maastricht
The NETHERLANDS 6200 MD

Ms. Carolyn R. Crone
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. C. M. Dayton
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 4112
University of Maryland
College Park, MD 20742

Dr. Lou DiBello
CERL
University of Illinois
103 South Mathews Avenue
Urbana, IL 61801

Dr. Dattprasad Divgi
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Mr. Hei-Ki Dong
Bell Communications Research
Room PYA-1K207
P.O. Box 1320
Piscataway, NJ 08855-1320

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Stephen Dunbar
224B Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. James A. Earles
Air Force Human Resources Lab
Brooks AFB, TX 78235

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Englehard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

Dr. Benjamin A. Fairbank
Operational Technologies Corp.
5825 Callaghan, Suite 225
San Antonio, TX 78228

Dr. P-A. Federico
Code 51
NPRDC
San Diego, CA 92152-6800

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Drew Gitomer
Educational Testing Service
Princeton, NJ 08541

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Michael Habon
DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Services
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Dr. Douglas H. Jones
1280 Woodfern Court
Toms River, NJ 08753

Dr. Brian Junker
Carnegie-Mellon University
Department of Statistics
Schenley Park
Pittsburgh, PA 15213

Dr. Milton S. Katz
European Science Coordination
Office
U.S. Army Research Institute
Box 65
FPO New York 09510-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State
University
P.O. Box 522
Murfreesboro, TN 37132

Mr. Soon-Hoon Kim
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. Richard J. Koubek
Department of Biomedical
& Human Factors
139 Engineering & Math Bldg.
Wright State University
Dayton, OH 45435

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Rodney Lim
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Dr. Robert Lockman
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. Clessen J. Martin
Office of Chief of Naval
Operations (OP 13 F)
Navy Annex, Room 2832
Washington, DC 20350

Dr. James R. McBride
The Psychological Corporation
1250 Sixth Avenue
San Diego, CA 92101

Dr. Clarence C. McCormick
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541